In: SFB 100 "Elektronische Sprachforschung" (Hrsg., 1972): Aspekte der automatischen Lemmatisierung. Bericht 10-72. Linguistische Arbeiten (LA) 12, 4-10

Harald H. Zimmermann

Zur Konzeption der automatischen Lemmatisierung von Texten.

Aus den Darlegungen in LA 10 (bes. S. 3ff.) kann man schließen, dass es sich bei dem Vorhaben 'Automatische Lemmatisierung' im Grunde um das Teilproblem einer auf linguistischen Kriterien aufbauenden Textdokumentation handelt. Einen sinnvollen ersten Abschnitt bildet u.E. dabei die Frage der Lemmatisierung, d.h. die Rückführung von Flexionsformen (Wortformen) auf einen graphematischen Repräsentanten bestimmter gemeinsamer Merkmale, den Lemmanamen. Ein Beispiel dafür: die Flexionsformen LEDIG LEDIGES LEDIGER LEDIGEM LEDIGEN LEDIGE sollen alle etwa die Merkmale (Adjektiv; menschliches Attribut; ohne Ehepartner...) haben und sich nur durch einige syntagmatische (Kasus-) Spezifikationen unterscheiden. Sie lassen sich repräsentieren durch den 'Lemmanamen' LEDIG als der ausgewählten (definitorisch festgelegten) Flexionsform. Bei Adjektiven ist dies die prädikative Form im Positiv. Problematisch wird die Frage der Repräsentation bei Synonymen wie FAHRSTUHL, LIFT, AUFZUG[1]). Hier wird deutlich, dass der graphematische Repräsentant nur ein Ersatz für die explizite Notation der Merkmale selbst sein kann. Will man dieses Phänomen beschreiben, so setzt man entweder mehrere Repräsentanten an und bringt diese in eine äquivalente Beziehung zueinander (implizit geschieht dies bereits durch das Aufzählen der (gleichen) Merkmale, explizit kann es durch Querverweise (Adressen) dargestellt werden), oder man verfährt ähnlich wie bei Schreibvarianten (FRISEUR - FRISÖR), indem man willkürlich eine dieser 'Varianten' als Lemmanamen - als Namen für diese Menge - auswählt.

> 1) Hier soll nicht über das generelle Problem der Synonymie als der völligen Identität aller Merkmale gerechtet werden; es wird also abgesehen von Fragen wie regionale Verwendung, Mehrdeutigkeit einzelner Begriffe usw.

Die Ausführungen in LA 10 (insbesondere die Kodierungsanweisungen mit den dabei aufgeführten semantischen Markierungen) machen jedoch deutlich, dass die Realität, die in den Lexikoneintragungen offenbar wird, von den theoretischen Vorstellungen und Ansprüchen noch ziemlich entfernt ist. Die aufgeführten Merkmale reichen bei weitem nicht aus, alle Lexikoneinträge explizit zu differenzieren. Die Repräsentation von (vorwiegend semantischen) Merkmalen geschieht in vielen Fällen nicht durch die Kodierung, also die ausdrückliche Angabe der spezifischen Eigenschaften oder Relationen, sondern nur implizit durch den Lemmanamen selbst; d.h. es gibt viele Lemmata, die unterschiedliche Bedeutungen repräsentieren, sich im Lexikon aber weder durch syntaktische noch durch semantische Merkmale unterscheiden. So lassen sich z.B. die Antonyme LEDIG und VERHEIRATET anhand der Merkmale nicht gegeneinander abgrenzen, auch TISCH, FLUSS oder BAUM sind zumindest auf syntaktischsemantischer Ebene nicht zu differenzieren.

Anderseits gelingt es aber, einen überwiegenden Teil der graphematischen Mehrdeutigkeiten bereits mit einem bescheidenen Klassifikationssystem zu erfassen, d.h. im Lexikon anhand der Markierungen zu differenzieren. Die vorliegende Kategorisierung und Subkategorisierung erlaubt es, alle syntaktischen Homographen auf Wortklassenebene zu beschreiben und zu trennen (z.B.

BILLIGEN als Verb oder als Adjektiv-Flexionsform); ein Teil der semantisch mehrdeutigen Lemmata hat entweder auch verschiedenartige syntaktische Merkmale (z.B. wirkt die Valenz bei Verben wie STIMMEN unterscheidend; vgl. LA 10 S. 50) oder das einfache 'eigentlich' semantische Merkmalinventar (± abstrakt, ± belebt, ± menschlich, ± zählbar, ± kollektiv) reicht aus, die Polyseme wenigstens teilweise zu differenzieren (z.B. PFERD als Lebewesen von PFERD als Schachfigur zu trennen). Natürlich bleibt ein nicht zu vernachlässigender Rest von Mehrdeutigkeiten (wie SCHLOSS (auf dem Berg) vs. (an der Tür)), bei denen das Beschreibungsinventar nicht ausreicht (d.h.: beidemal wird SCHLOSS mit denselben Merkmalen ausgestattet).

Zwei Gründe sind dafür bestimmend, dass vorläufig auf den weiteren Ausbau der semantischen Komponente verzichtet wird, der allein zu weiteren Differenzierungen beitragen kann: Zunächst ist zu erwarten, dass die konkreten Ergebnisse einer automatischen Texterschließung auf der Basis der nach dem Saarbrücker Verfahren ermittelten Lemmata [1)] bereits einen wesentlichen Fortschritt gegenüber den bisherigen mechanischen Vorgehensweisen bringen, die sich überwiegend in der Herstellung einfacher Wortformenlisten (nach alphabetischer, rückläufiger oder Häufigkeits-Ordnung) erschöpfen.

> 1) Es sei noch einmal darauf hingewiesen, dass dazu eine automatische syntaktische Analyse durchgeführt wird, die auf den Merkmalen operiert. Vgl. S. 11 f

Die Zahl manueller Eingriffe [2)] in den Analyse- und Dokumentationsprozess wird in der wichtigen Frage der Mehrdeutigkeitsauflösung zumindest stark herabgesetzt, die Nachredaktion bzw. eine komplexe Präparierung von Texten zur maschinellen Analyse wird vermieden oder jedenfalls eingeschränkt. Dabei lässt sich der Umstand ausnutzen, dass die Lemmata als Repräsentanten der Merkmale gelten können; der Benutzer entsprechender Text-Wörterbücher hat dabei ebenfalls von den Repräsentanten auszugehen. Es ist allerdings auch möglich, die im Lexikon bereits verzeichneten Informationen abzufragen, d.h. Textinformationen danach zu ordnen oder auszuwählen: eine Erweiterung dieser Basis - vor allem der semantischen Komponente (man denke etwa an hierarchische Strukturierungen, also die Über- oder Nebenordnung von Lemmata) - ließe allerdings im Hinblick auf satz- und textsemantische Analysen weitere Verbesserungen erwarten.

> 2) Es leuchtet wohl ein, dass auf der Basis dieses Merkmalsystems noch keine `produktive` vollautomatische Lemmatisierung möglich sein kann. Daher sind zu diesem Zweck im Rahmen eines interaktiven Systems Benutzereingriffe vorgesehen, die vor allem die Auflösung von maschinell nicht lösbarer. Mehrdeutigkeiten zum Ziel haben.

Diese an sich erstrebenswerte umfassendere Konzeption einer Textanalyse ist in naher Zukunft nicht zu erreichen.

Dies um so weniger, als dazu kaum praktische Vorarbeiten (also abgesehen von theoretischen Ansätzen) geleistet sind [1)]. So fehlte bisher ein maschinell zugängliches umfangreiches Wörterbuch zur deutschen Sprache; das einzige seit einigen Jahren benutzbare und benutzte Maschinenlexikon, aufbauend auf dem Wörterbuch von Mackensen, liefert über den Wortlaut hinaus zu wenig Informationen [2)], als dass es sinnvoll im Rahmen einer automatischen Sprachanalyse verwendet werden könnte. Im Vordergrund der Saarbrücker Arbeiten steht daher zunächst die Herstellung eines geeigneten maschinellen Lexikons; es soll also möglichst so beschaffen sein, dass die Ergebnisse einer ersten Kodierphase eine hinreichende Grundlage zur automatischen

Lemmatisierung bilden und zugleich eine spätere Erweiterung - vor allem im morphologischen und semantischen Bereich - erleichtern.

1) Ein praktisches Modell stellen etwa die Versuche von Quillian dar: Vgl. M.Minsky (ed.): Semantic Information Processing, Cambridge Mass. (MIT-Press) , 1968; darin: R.M. Quillian, Semantic Memory.
2) Notiert werden etwa noch die Wortklassenangaben und bei Nomina das Genus.

Die Konzeption eines Lexikons schließt die Konzeption einer Grammatik weitgehend ein: Lexikon und Regelsystem bilden eine Einheit. Bereits in LA 10 (bes. S. 11ff.) wurde darauf hingewiesen, dass eine Satzanalyse (oder weiter gefasst: eine Kontextanalyse) erst die Voraussetzung dafür schafft, Texte zu lemmatisieren. Die an der (Wort- oder Satz-) Oberfläche mehrdeutigen (Teil-) Strukturen sind mittels der Informationen aus dem Kontext zu vereindeutigen und in den Rahmen der Strukturierung des Textes (oder bescheidener: der Sätze) entsprechend einzugliedern. Oberflächengrammatiken - wie sie auch dem Saarbrücker Analysemodell von 1969 zugrundelagen - sind eine erste Voraussetzung, aber für eine Strukturbeschreibung nicht ausreichend. Daraus folgt, dass dem Analysealgorithmus eine adäquatere Grammatik zugrunde gelegt werden muss. Die dafür in Grundzügen entwickelte Grammatik ist im Kern an dem generativ-transformationellen Modell orientiert [1)] . Im Hinblick auf die automatische Lemmatisierung dient die Analyse (sie ist zunächst als Satzanalyse konzipiert [2)] , d.h. es werden nur Informationen benutzt und erarbeitet, die innerhalb des Satzrahmens verfügbar sind) der Auflösung von Mehrdeutigkeiten. Dass ihr im Rahmen einer Textdokumentation weitere Funktionen zukommen können, erhöht ihren Wert.

1) Eine Beschreibung des Regelsystems wird in einem reiferen Arbeitsbericht gegeben.
2) Sie ist die Voraussetzung einer umfassenderen Textanalyse.

Einige entscheidende Veränderungen gegenüber dem Satz-analysemodell von 1969[3)] wird auch der Analysealgorithmus erfahren: Zwar kann gegenwärtig das Verfahren noch nicht in Einzelheiten übersehen werden - die bisherigen Vorarbeiten haben etwa noch nicht klar erkennen lassen, inwieweit das prinzipiell erstrebenswerte Ziel, Grammatik (sprachliches Regelsystem) und Parser (maschineller Algorithmus) zu trennen, in allen Teilen realisierbar ist, ohne dass die Speicherkapazität überschritten [4)] bzw. die Rechenzeiten für die Satzanalyse unvertretbar groß werden. Dennoch wird von folgenden Prinzipien auszugehen sein:

3) Vgl. H. Eggers et al., Elektronische Syntaxanalyse
4) Sie beträgt bei dem verfügbaren Computer maximal 60 K Worte (à 24 Bits) für Instruktionen, wenn ohne komplizierte und zeitraubende Overlay-Techniken gerechnet werden soll.

a) Alle aufgrund der zugrundegelegten Grammatik möglichen Strukturen werden ermittelt, d.h. im wesentlichen: alle aufgrund des Regelsystems (noch) mehrdeutigen Satz- oder Teil-Strukturen werden angegeben.
b) Der Erkennung von Eigennamen und anderen, nicht im Lexiken verzeichneten oder über Lexikoneinträge erschließbaren, d.h. 'unbekannten' Wörtern wird Rechnung getragen (über Wortbildungs- und Kontextregeln).
c) Durch die Integration eines interaktiven Kommunikationssystems (Mensch-Maschine) soll die Möglichkeit gegeben werden, in Zweifelsfällen auf programmierte Anfragen des Computers zu reagieren und damit bei maschinell schwer entscheidbaren Problemen so weit Hilfestellung zu

geben, dass die weitere automatische Analyse und damit die Lemmatisierung (in einer produktiven Phase) zu brauchbaren Ergebnissen führt.

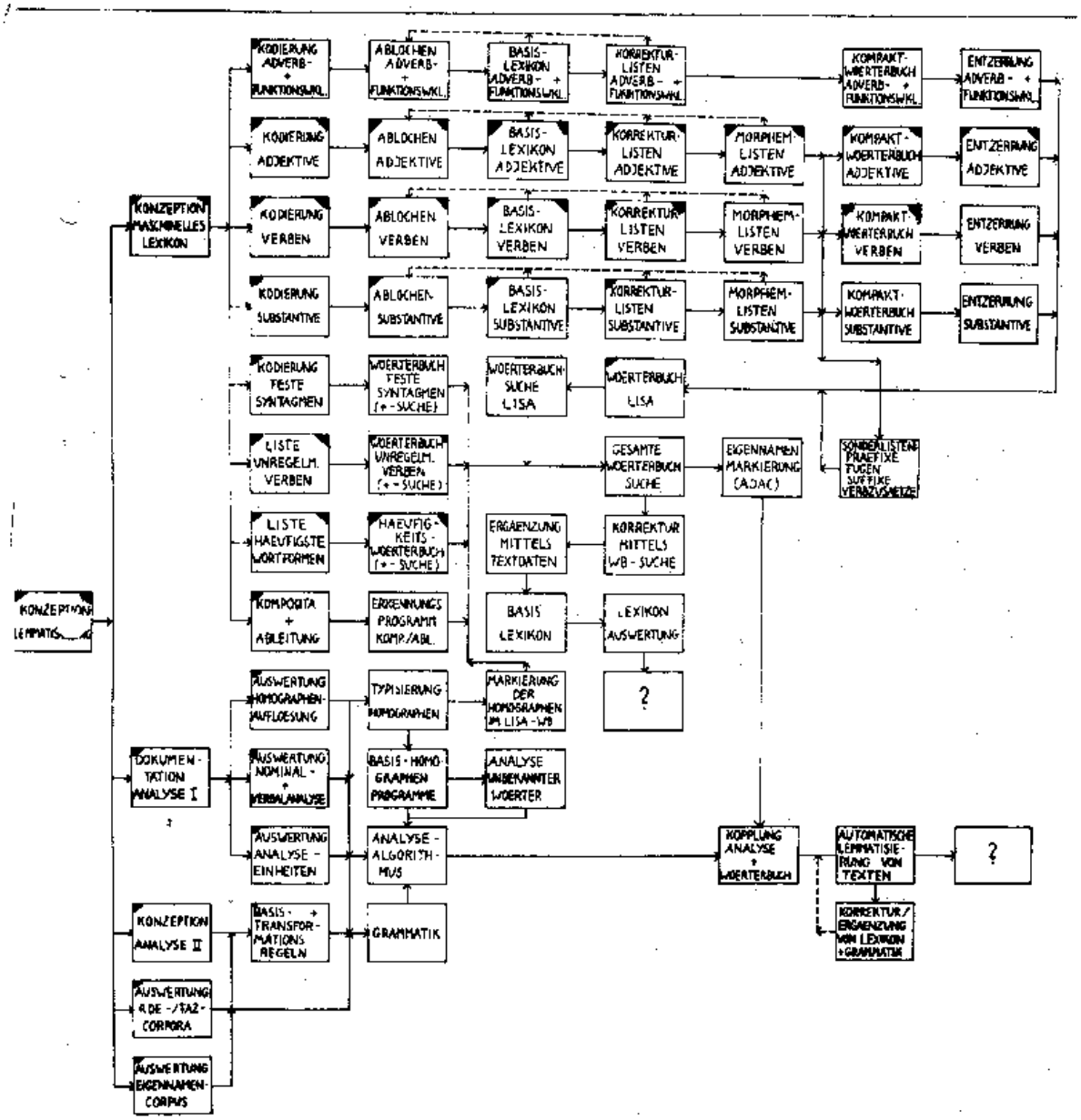
Bei der Konzeption des Lemmatisierungsverfahrens sollen Erfahrungen und Ergebnisse bisheriger Arbeiten der Saarbrücker Forschungsgruppe in weitem Maße berücksichtigt werden. Das Grammatikmodell ist dabei im Vergleich mit Auswertungen des Saarbrücker Textcorpus (rd. 11.000 nach syntaktischen Gesichtspunkten klassifizierte Sätze) zu überprüfen; eine ins Einzelne gehende Beschreibung des Analyseverfahrens von 1969 [1)] wird beim Aufbau des Parsers berücksichtigt werden.

1) Diese teilweise in Flussdiagrammform dargestellte Dokumentation wird voraussichtlich im Herbst 72 fertiggestellt sein. Interessenten werden um entsprechende Mitteilung gebeten.

Die folgende Übersicht über die Abfolge der Arbeiten fasst noch einmal zusammen, wie sich das Konzept der automatischen Lemmatisierung gegenwärtig darstellt. Die stärkere Differenzierung des Lexikonteils [2)] zeigt, dass dieser Bereich gegenwärtig im Mittelpunkt der Realisierung steht.

2) Vgl. dazu den Artikel 'Das Lexikonsystem zur maschinellen Sprachbearbeitung' in diesem Bericht.

Abb. 1972_a_1


KONZEPTION MASCHINELLES LEXIKON
KODIERUNG ADVERB- + FUNKTIONSWK.
ABLOCHEN ADVERB- + FUNKTIONSWK.
BASIS-LEXIKON ADVERB- + FUNKTIONSWK.
KORREKTUR-LISTEN ADVERB- + FUNKTIONSWK.
KOMPAKT-WOERTERBUCH ADVERB- + FUNKTIONSWK.
ENTZERRUNG ADVERB- + FUNKTIONSWK.
KODIERUNG ADJEKTIVE
ABLOCHEN ADJEKTIVE
BASIS-LEXIKON ADJEKTIVE
KORREKTUR-LISTEN ADJEKTIVE
MORPHEM-LISTEN ADJEKTIVE
KOMPAKT-WOERTERBUCH ADJEKTIVE
ENTZERRUNG ADJEKTIVE
KODIERUNG VERBEN
ABLOCHEN VERBEN
BASIS-LEXIKON VERBEN
KORREKTUR-LISTEN VERBEN
MORPHEM-LISTEN VERBEN
KOMPAKT-WOERTERBUCH VERBEN
ENTZERRUNG VERBEN
KODIERUNG SUBSTANTIVE
ABLOCHEN SUBSTANTIVE
BASIS-LEXIKON SUBSTANTIVE
KORREKTUR-LISTEN SUBSTANTIVE
MORPHEM-LISTEN SUBSTANTIVE
KOMPAKT-WOERTERBUCH SUBSTANTIVE
ENTZERRUNG SUBSTANTIVE
KODIERUNG FESTE SYNTAGMEN
WOERTERBUCH FESTE SYNTAGMEN (+-SUCHE)
WOERTERBUCH-SUCHE LISA
WOERTERBUCH LISA
LISTE UNREGELM. VERBEN
WOERTERBUCH UNREGELM. VERBEN (+-SUCHE)
GESAMTE WOERTERBUCH SUCHE
EIGENNAMEN MARKIERUNG (ADAC)
SONDERLISTEN PRAEFIXE FUGEN SUFFIXE VERBZUSAETZE
LISTE HAEUFIGSTE WORTFORMEN
HAEUFIGKEITS-WOERTERBUCH (+-SUCHE)
ERGAENZUNG MITTELS TEXTDATEN
KORREKTUR MITTELS WB-SUCHE
KOMPOSITA + ABLEITUNG
ERKENNUNGS PROGRAMM KOMP./ABL.
BASIS LEXIKON
LEXIKON AUSWERTUNG
?
AUSWERTUNG HOMOGRAPHEN-AUFLOESUNG
TYPISIERUNG HOMOGRAPHEN
MARKIERUNG DER HOMOGRAPHEN IM LISA-WB
DOKUMENTATION ANALYSE I
AUSWERTUNG NOMINAL- + VERBALANALYSE
BASIS-HOMOGRAPHEN PROGRAMME
ANALYSE UNBEKANNTER WOERTER
AUSWERTUNG ANALYSE-EINHEITEN
ANALYSE-ALGORITHMUS
KOPPLUNG ANALYSE + WOERTERBUCH
AUTOMATISCHE LEMMATISIERUNG VON TEXTEN
?
KONZEPTION ANALYSE II
BASIS- + TRANSFORMATIONS REGELN
GRAMMATIK
KORREKTUR/ERGAENZUNG VON LEXIKON + GRAMMATIK
AUSWERTUNG DE-/FR-/EN-CORPORA
AUSWERTUNG EIGENNAMEN-CORPUS